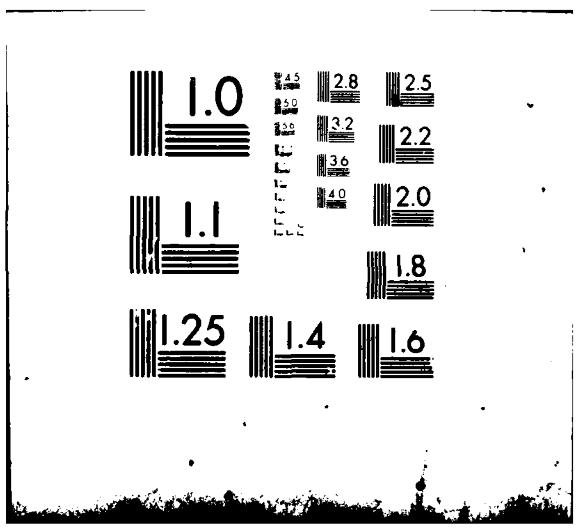
AD-A084 041   DESMATICS INC STATE COLLEGE PA   F/G 14/2
AN INVESTIGATION OF A STATISTICAL PROCEDURE FOR MONITORING TWO---ETC(U)
APR 80 L A KALISH, D E SMITH   N00014-75-C-1054
UNCLASSIFIED TR-106-9   NL

1 of 1
AD
A084041

END
DATE
FILMED
6-80
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

DESMATICS, INC.

P. O. Box 618
State College, Pa. 16801
Phone: (814) 238-9621

Applied Research in Statistics - Mathematics - Operations Research

AN INVESTIGATION OF A STATISTICAL PROCEDURE FOR MONITORING TWO-SAMPLE LIFE TESTS.

by

Leslie A. Kalish
and
Dennis E. Smith

TECHNICAL REPORT NO. 106-9

April 1980

This study was supported by the Office of Naval Research under Contract No. N00014-75-C-1054 Task No. NR 042-334.

ABSTRACT

A statistical procedure is described for comparing two lifetime distributions when the data is reviewed repeatedly over time. The procedure provides the capability of early decision while maintaining both a fixed significance level and a fixed maximum length for the entire experiment. The effects which staggered entry, number of looks at the data and maximum test length have on power and expected test length are discussed. An application is made to a bearing fatigue-life test.

Key Words:

Life Tests

Repeated Testing

Staggered Entry

Power

Expected Test Length

TABLE OF CONTENTS

I. INTRODUCTION

In the field of reliability, it is often necessary to conduct experiments in order to compare two lifetime distributions. Examples include comparing lifetimes of similar products from different manufacturers or comparing similar products which meet different quality control standards. Due to the length of such "life tests," it is usually desirable to review the data periodically over the course of the experiment. If the difference between the two groups is large, it may be possible to reach an early conclusion and truncate the test with considerable savings in time and expense.

When data is reviewed periodically, the probabilities associated with classical significance tests are not appropriate. Wald [3] developed sequential procedures wherein the data is reviewed periodically with a fixed significance level for the entire test. The Wald type sequential tests are useful primarily when the response time is short relative to the entry period of items into the experiment [1]. In the area of reliability, response time (for example, fatigue life of bearings) can be several years.

Canner [1] developed a statistical procedure for relatively short tests, which is a compromise between the fixed length and the sequential tests. The "Canner test" provides the capability of early decision while maintaining both a fixed significance level and a fixed maximum length for the entire experiment. Items can enter the experiment simultaneously or with a staggered entry following the uniform distribution. (Actually any entry distribution can be assumed, but only the simultaneous and

uniform cases are considered here.)

Although the Canner test was originally developed for the analysis of clinical trials, there is no reason to limit its use to medical applications. In this report, the Canner test is described in a reliability setting. After some of Canner's results are mentioned, additional properties of the test are investigated. In particular, the power and expected test length are approximated for a variety of conditions and the effects which various experimental design factors have on power and expected test length are discussed.

II. THE CANNER TEST

The control group, group 1, and the experimental group, group 2, contain n_1 and n_2 items respectively. Let T_{ij} denote the times of entry into the experiment, assumed to be uniformly distributed over the first T_E years, let R_{ij} denote the failure times (in years) and let T_F denote the maximum length of the test (in years). Here, as in the remainder of the report, i indicates group and j indicates item ($i = 1, 2; j = 1, 2, \ldots, n_i$). For example, T_{ij} denotes the entry time of the $j^{\underline{th}}$ item from the $i^{\underline{th}}$ group. For ease of writing, this item will be referred to as the $(i, j)^{\underline{th}}$ item.

The response variable is item lifetime, which is assumed to follow a Weibull distribution with known shape parameter, ν. Although Canner [1] considers only the exponential distribution, which is Weibull with $\nu = 1$, the more general case can be considered by utilizing the fact that if X is distributed as a Weibull random variable with scale parameter γ and shape parameter ν, then X^ν is distributed as an exponential random variable with mean γ^ν.

At time t (years), $0 \le t \le T_F$, the $(i,j)^{\underline{th}}$ item is in one of three states:

1. the item failed at time R_{ij} $(T_{ij} \le R_{ij} \le t)$
2. the item remains in operation $(T_{ij} \le t \le R_{ij})$
3. the item has not yet entered the test $(t \le T_{ij} \le R_{ij})$.

For the n_{it} items from group i which are in states 1 or 2 at time t,

define u_{ijt}, the survival time in "exponential" units, as

$$u_{ijt} = \begin{cases} (R_{ij} - T_{ij})^{\nu}, \text{ if the } (i, j)^{\underline{th}} \text{ item is in state 1 at time t.} \\ (t - T_{ij})^{\nu}, \text{ if the } (i, j)^{\underline{th}} \text{ item is in state 2 at time t.} \end{cases}$$

Let $$\delta_{ijt} = \begin{cases} 1, \text{ if the } (i, j)^{\underline{th}} \text{ item is in state 1 at time t} \\ 0, \text{ if the } (i, j)^{\underline{th}} \text{ item is in state 2 at time t,} \end{cases}$$

and define $D_{it} = \sum_{j=1}^{n_{it}} \delta_{ijt}$ and $U_{it} = \sum_{j=1}^{n_{it}} u_{ijt}$. Note that at time t, u_{ijt} denotes the length of time (after transforming to exponential units) that the $(i, j)^{\underline{th}}$ item has been in operation, n_{it} denotes the number of items from group i which have entered the experiment, D_{it} denotes the number of items from group i which have failed and U_{it} denotes the accumulated, transformed time for all items in group i.

The probability that an item in group i fails before t months of operation time is $P_i(t) = 1 - e^{-\lambda_i t^{\nu}}$, where $\lambda_i = \gamma_i^{\nu}$ is the exponential parameter for group i. The purpose of the test is to determine whether λ_1 and λ_2 differ. Throughout this report the one-sided test H_0: $\lambda_1 = \lambda_2$ versus H_1: $\lambda_1 < \lambda_2$ will be considered. Extensions to the test with the inequality reversed or to the two-sided test are straightforward. See Canner [1] for construction of the two-sided test.

The maximum likelihood estimate of $\lambda_1 - \lambda_2$ is given by

$$\hat{\lambda}_1 - \hat{\lambda}_2 = \frac{D_{1t}}{U_{1t}} - \frac{D_{2t}}{U_{2t}} .$$

An estimator of the variance of $\hat{\lambda}_1 - \hat{\lambda}_2$ is given by

$$\hat{Var}(\hat{\lambda}_1 - \hat{\lambda}_2) = \hat{\lambda}^2 \left[\frac{n_{1t} + n_{2t}}{D_{1t} + D_{2t}}\right] \left[\frac{1}{n_{1t}} + \frac{1}{n_{2t}}\right] ,$$

where $\hat{\lambda} = (D_{1t} + D_{2t})/(U_{1t} + U_{2t})$.

It is decided in advance that the data will be reviewed K times during the course of the experiment, at intervals of T_F/K years. For the k^{th} look at the data (at time $t = kT_F/K$), the following test statistic is computed:

$$z_k = (\hat{\lambda}_1 - \hat{\lambda}_2)/\sqrt{\hat{Var}\ (\hat{\lambda}_1 - \hat{\lambda}_2)} \quad .$$

The decision rule is

1. If $z_k \leq A$, stop the experiment and conclude H_1.
2. If $z_k > A$,
   a. continue the experiment if $k < K$
   b. stop the experiment and conclude H_0 if $k = K$.

A is the critical value of the test such that $P(z_k \leq A|H_0) = \alpha$ for a given level of significance, α.

Due to the flexibility of the Canner test, analytic construction of the exact test quickly becomes intractable. Therefore, a computer simulation is used to construct an approximate test for a given set of restrictions: sample size (n_1, n_2), maximum length of test (T_F), number of looks at the data (K), length of entry period (T_E), value of Weibull scale parameter under the null hypothesis of no difference between group 1 and group 2 ($\gamma_1 = \gamma_2$), value of the Weibull shape parameter (ν), significance level (α). Following is an outline of the simulation procedure.

1. Specify n_1, n_2, T_F, K, T_E, $\gamma_1 (= \gamma_2)$, ν, α.
2. Simulate entry times T_{ij} by generating independent, uniformly distributed random numbers between 0 and T_E. For the case of simultaneous entry, $T_E = 0$ and $T_{ij} = 0$ ($i = 1, 2$; $j = 1,\ldots,n_i$).
3. Generate a new set of independent uniform random numbers between

0 and 1 and use these to simulate failure times in accordance with the failure distribution under the null hypothesis as follows. Suppose that for the (i, j)$\underline{\text{th}}$ item, the random number x is chosen. Then $x = 1 - e^{-\gamma_i^\nu y^\nu}$, which implies that the operation time until failure is $y = [-\ln(1 - x)]^{1/\nu}/\gamma_i$. Thus the simulated failure time is $R_{ij} = T_{ij} + y = T_{ij} + [-\ln(1 - x)]^{1/\nu}/\gamma_i$. If $R_{ij} > T_F$, that item is considered to have survived the maximum length of the experiment without failure.

4. Calculate z_k $(k = 1,\dots,K)$ using the values simulated in steps 2 and 3.
5. Define $z = \min\limits_1^K z_k$.
6. Repeat steps 2 to 5, m times. The lower 100 α percentile of the m values of z will define the appropriate critical value, A. For the special case of the symmetric distribution of z_k resulting from equal sample sizes $(n_1 = n_2)$, the critical value is defined by the lower $100 \cdot 2 \cdot \alpha$ percentile of the m values of $-|z|$. For this report, the value of m was 5000 for each simulation of a critical value.

III. BEHAVIOR OF THE TEST

By simulating an experiment under various sets of restrictions, much can be learned about the behavior of the test. The effects of changing restrictions on the critical region are given by Canner [1]. Once a critical region is constructed for a given set of restrictions, the experiment can be simulated with only the value of γ_2 changed (i.e., under the alternative hypothesis) to estimate the power of the test and the expected test length. The proportion of the m values of z which fall in the critical region is the estimate of power. In addition, the expected test length is estimated by averaging the m simulated times until decision. For this report, the value of m was 3000 for each simulation of power and expected test length.

A. ACCURACY OF POWER ESTIMATES

Before powers can be meaningfully compared it is necessary to determine how accurate the power approximations are. Canner [1] deals with part of this question indirectly by constructing 95% confidence intervals for critical values using a nonparametric method given by Kendall and Stuart [2]. By estimating the power of a test using the upper and lower 95% confidence limits on the critical value (as well as the point estimate of the critical value), a "narrow" confidence interval can be calculated for the power estimate. The qualification, "narrow," is used because the additional component of variability due to the run to run differences of the power simulation is absent.

To estimate this component, a power simulation can be replicated several times with only different random number streams, and a 95% confidence interval

constructed from the power estimates. The sum of the width of this interval and the width of the "narrow" interval mentioned earlier gives the approximate width of the true 95% confidence interval. Ideally, this procedure should be followed for each set of restrictions considered. However, since only a rough idea of the magnitude of error is usually needed, just a few representative cases were used. From these, it was concluded that the power estimates are accurate to about $\pm$ .027, 95% of the time. It should be stressed that this is nothing more than a rough rule of thumb.

B. CANNER'S RESULTS

Some of Canner's results [1] deal with the effects that varying sample size, scale parameter and number of looks at the data have on the critical region. Cases with both simultaneous and staggered entry are considered. He also discusses the effect that increasing the number of looks has on the power of the test, noting a moderate loss in power as the number of looks increases.

In an investigation of robustness, Canner shows the test to be quite robust against changes in entry distribution and length of test. As it is not the purpose of this report to give a complete review of Canner's work, the reader is referred to Canner's paper [1] for more results and details.

C. EXAMPLE OF AN EXPERIMENT

For the remainder of this section, an example experiment will be utilized. This example is based on a bearing fatigue-life test being considered by the Naval Ship Research and Development Center. The control group consists of bearings which are expected to have B_{10} lives of at least 10,000 hours. A

bearing is said to have a B_{10} life of h hours if the bearing fails within h hours of operation time with probability .10. The purpose of the study is to test whether bearings of the type in the experimental group have B_{10} lives significantly less than 10,000 hours. The sample size in each group is 50.

Bearings in both groups are assumed to have Weibull fatigue-life distributions with shape parameter 1.5 and scale parameters which are functions of the B_{10} lives. It is convenient to express the hypotheses in terms of B_{10} lives rather than scale parameters. Letting B_A and B_B denote the B_{10} lives, in thousands of hours, of the control group and the treatment group respectively, the hypotheses can be written

$$H_0: \; B_A = 10, \; B_B = 10$$
$$H_1: \; B_A = 10, \; B_B < 10.$$

For the power calculations which follow (except where noted otherwise) the value of B_B under the alternative hypothesis is assumed to be 5.0.

D. RESULTS CONCERNING POWER OF THE TEST

An alternative to Canner's test, which can be used when all items enter the test simultaneously and the data is reviewed only once at the end of the experiment, is to invoke the Central Limit Theorem and compare the average censored lifetimes of the two groups. The test used is essentially the standard two-sample t-test. Letting

$$x_{ij} = \begin{cases} R_{ij}, & \text{if } R_{ij} < T_F \\ T_F, & \text{if } R_{ij} \geq T_F, \end{cases}$$

$\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$, and $s_i^2 = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2/(n_i - 1)$, the test statistic is

$$W = (\bar{x}_1 - \bar{x}_2)/[(s_1^2/n_1) + (s_2^2/n_2)]^{1/2} .$$

Under H_0 and for large n_1 and n_2, W has essentially a standard normal distribution. Note that for this simple case ($T_E = 0$, $K = 1$) the major difference between the t-test and the Canner test is the test statistic. A 1.25 year experiment was simulated to compare the powers of the t-test and the Canner test. The results, on lines 1 and 2 of Figure 1, show that the power of the Canner test is better than that of the t-test.

A second, more realistic experiment was simulated to illustrate the effects that staggered entry and repeated looks at the data have on power and and expected test length. In this experiment, an entry period of 2.5 years ($T_E = 2.5$) and a maximum length of 2.5 years ($T_F = 2.5$) were assumed. The powers and expected test lengths were estimated for both $K = 1$ and for $K = 10$. The results, on lines 3 and 4 of Figure 1, show that when repeated looks are made at the data, a moderate loss in power is offset by a significant decrease in expected test length. (For the case $K = 1$, expected test length is, of course, exactly the full length of the test, T_F.)

Another interesting comparison can be made from lines 2 and 3. Note that in the experiment represented by line 3, items enter over the whole course of the test so that the average follow-up time is .5 T_F or 1.25 years. Thus lines 2 and 3 represent experiments with equal average follow-up times. When average follow-up time is held fixed and one look at the data is made, staggered entry appears to increase the power slightly. It might be argued that due to the error bounds of $\pm$ .027 around each power estimate, there is no significant difference between the powers of the staggered and simultaneous entry cases.

Line	Number of Looks (K)	Entry Period in Years (T_E)	Maximum Test Length in Years (T_F)	Average Follow-up Time in Years	Power	Expected Test Length in Years
1	1	0.0	1.25	1.25	.746	1.25
2	1	0.0	1.25	1.25	.835	1.25
3	1	2.5	2.5	1.25	.844	2.5
4	10	2.5	2.5	1.25	.779	1.88
5	5	0.0	1.25	1.25	.751	.87
6	6	0.0	6.0	6.0	.853*	3.25*
7	12	12.0	12.0	6.0	.759*	8.15*

Figure 1. Power and Expected Test Length for Various Tests. Calculations are based on a sample size of 50 per group, significance level .10, $B_B = 5.0$ (Except $B_B = 6.7$ where (*) appears.) See Text.

However, while the difference for this particular case is quite small, it does exist; an intuitive argument will be given in section III.E.

In the experiment represented by line 4 of Figure 1, the items have a staggered entry and the data is reviewed quarterly (every three months) with an average follow-up time of 1.25 years. Consider the experiment represented by line 5, also with quarterly review and average follow-up time of 1.25 years, but with simultaneous entry. Referring to lines 4 and 5, it is seen that there is a slight increase in power due to the staggered entry when the period of data review and average follow-up time are held fixed.

Note that in the last two examples the effect of staggered entry has been to increase the power. However, for cases where the test length is relatively long compared to item lifetimes (cases not likely to be encountered in practice) staggered entry can have a detrimental effect on power; an intuitive explanation for this will be offered in section III.E.

Consider two such "long" experiments, each with an average follow-up time of 6 years and with yearly review of the data. In one experiment, $T_E = 0$ and $T_F = 6$ while in the other experiment, $T_E = 12$ and $T_F = 12$. For these experiments, $B_B = 6.7$ is used in the alternative hypothesis instead of $B_B = 5.0$ so that the resulting powers have easily comparable values. Here, the effect of staggered entry (as is shown on lines 6 and 7) is to decrease the power.

E. EFFECTS OF STAGGERED ENTRY

The results given above are not surprising, except perhaps for the difference in the effect of staggered entry for short and long tests. When the underlying distribution is in fact exponential (i.e., Weibull with $\nu = 1.0$) or Weibull with $0 < \nu < 1$, this difference does not occur; the effect of

staggered entry then is always to decrease the power. For Weibull distributions with $\nu > 1$, the detrimental effect of staggered entry on power occurs only for "long" tests. In this case, the detrimental effect occurs only for longer and longer tests as ν increases. Recall that in the examples used above $\nu = 1.5$. The following paragraphs provide an intuitive explanation of this phenomenon.

In general, an uncensored observation (knowing the actual lifetime of an item) gives more information and yields better estimates and more powerful tests than a censored observation (knowing only that the lifetime of the item exceeds some fixed time). Thus, as the probability of failure before termination of a test increases, so does the power of the test.

Let $F(t)$ represent the cumulative distribution function of an underlying fatigue-failure probability distribution. $F(t)$ is the probability that an item fails before t years of operation time. Now consider an extreme type of staggered entry wherein half of the items enter the test at time 0 and the other half enter at time t'. If the test is terminated at time t'' ($t'' > t'$), then half of the items will have follow-up time $t_1 = t'' - t'$ and half of the items will have follow-up time $t_2 = t''$. The average follow-up time is $(t_1 + t_2)/2$ and the average probability of failure before termination of the test is $F_{stag} = [F(t_1) + F(t_2)]/2$. For the corresponding simultaneous entry test with the same follow-up time, the average probability of failure before termination of the test is $F_{sim} = F[(t_1 + t_2)/2]$. Thus, staggered entry will increase power when $F_{stag} > F_{sim}$ and will decrease power when $F_{stag} < F_{sim}$.

Figures 2 and 3 illustrate how the effect of staggered entry can vary, depending on the length of the test and the underlying distribution. In Figure 2, the cumulative distribution function of an exponential distribution

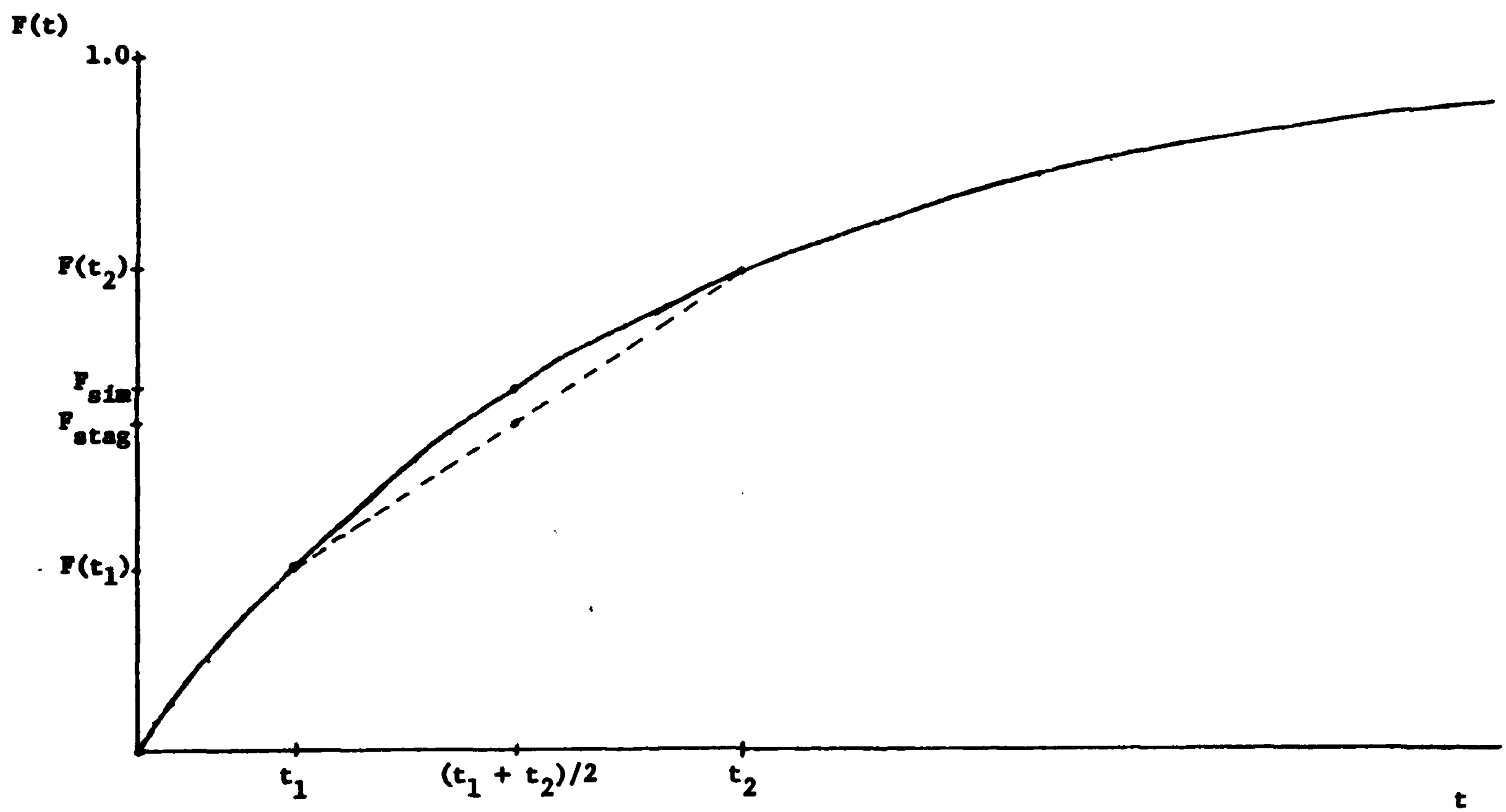


Figure 2. Cumulative Distribution Function of Exponential Distribution

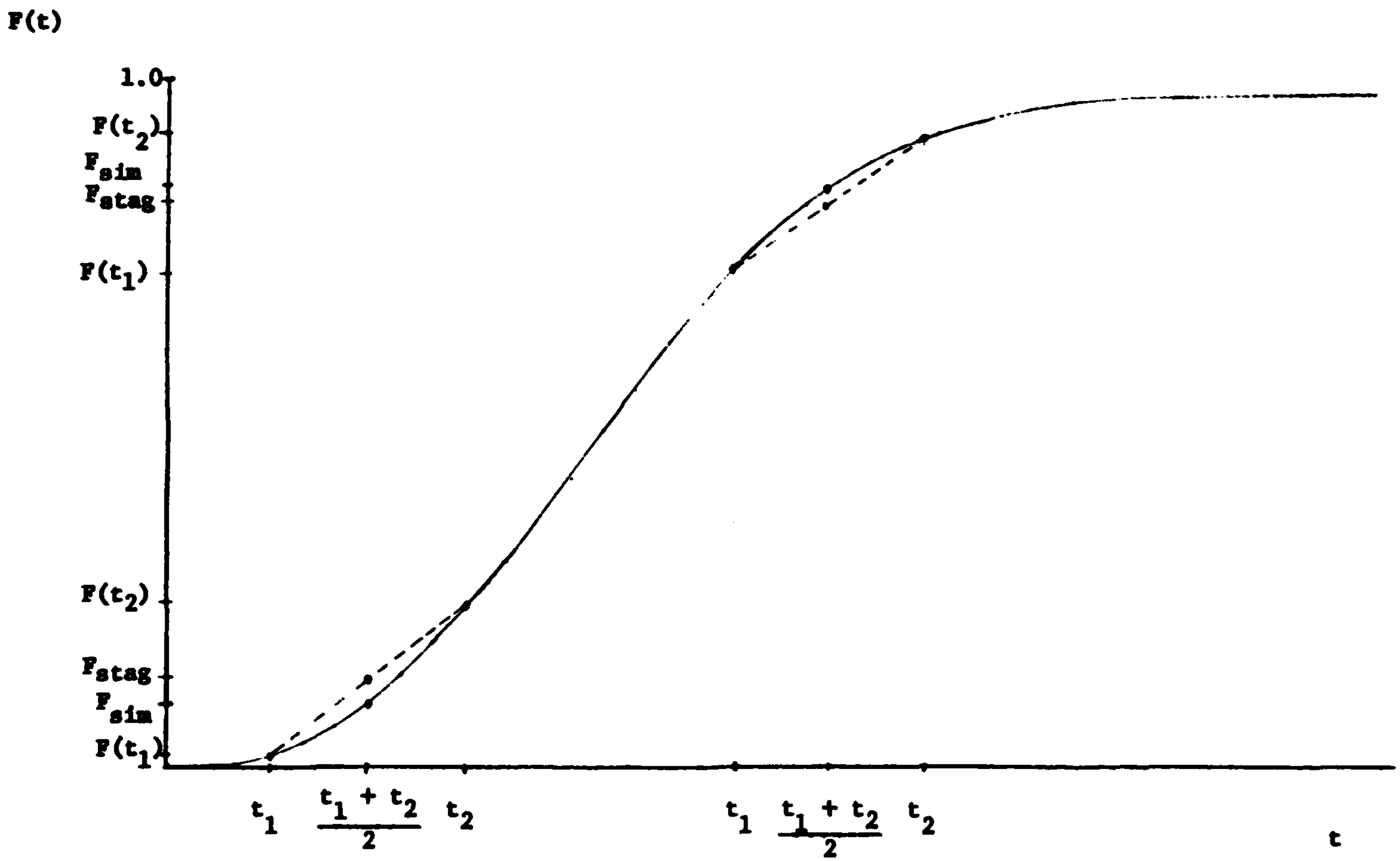


Figure 3. Cumulative Distribution Function of Weibull Distribution with Shape Parameter $\nu = 2.5$.

is shown. It can be seen that due to the concave shape of the function, $F_{stag} \leq F_{sim}$ for any values of t_1 and t_2. The same phenomenon occurs for any Weibull distribution with $0 < \nu \leq 1.0$. In Figure 3, the cumulative distribution function of the Weibull distribution with shape parameter $\nu = 2.5$ is shown. (Any $\nu > 1$ could have been used.) Note that for the larger values of t_1 and t_2, $F_{stag} < F_{sim}$ while for the smaller values of t_1 and t_2 the direction of the inequality is reversed. Thus, staggered entry can have a positive effect on power for short tests with a fixed average follow-up time.

When average follow-up time is fixed, it is important to note that, from a practical standpoint, the effect of staggered entry on power is largely a moot point. If an experimenter does have the opportunity to choose between different entry distributions and test lengths, it is always best to extend the average follow-up time as long as possible by using simultaneous entry and a long maximum test length. However, even if entry distribution and maximum test length are fixed (thus fixing an average follow-up time) it may be useful to estimate how much power is "lost" due to a staggered entry.

IV. DISCUSSION AND SUMMARY

While it is useful to exhibit the effects that entry time, maximum test length, number of looks and other design factors have on power with actual (simulated) examples, it must be remembered that it is often difficult to use these few examples to formulate hard and fast rules. A complete analysis of the interacting effects of various factors is not within the scope of this report.

However, the following statements about the Canner test can be made with high confidence. Some of the statements are drawn from well known results in the area of hypothesis testing while others are drawn from the results given in Canner's paper [1] and in section III of this report.

1. The Canner test is more powerful than the t-test.
2. Power increases as
   (a) maximum length of test increases.
   (b) true B_{10} life of the experimental group decreases.
   (c) number of looks at the data decreases.
3. Expected test length decreases as
   (a) number of looks at the data increases.
   (b) true B_{10} life of the experimental group decreases.
4. When average follow-up time is fixed and the underlying distribution is Weibull with $0 < \nu \leq 1.0$, staggered entry decreases the power.
5. When average follow-up time is fixed and the underlying distribution is Weibull with $\nu > 1$,
   (a) staggered entry increases the power for "short" tests.

(b) staggered entry decreases the power for "long" tests.

In real-life applications, even in the simplest of cases, it is likely that the experimenter would want to run several simulations to approximate the power of various tests as an aid in designing the experiment. At such a time, the nature of the interactions between maximum test length, entry time, number of looks and other design factors can be determined over a range of designs which are within economic and time constraints. In addition, the relative importance of power and expected test length can be considered as part of the design process.

V. REFERENCES

[1] Canner, P. L., "Monitoring Treatment Differences in Long-Term Clinical Trials," Biometrics, Vol. 33, pp. 603-615, 1977.

[2] Kendall, M. G. and Stuart, A., The Advanced Theory of Statistics, Vol. 2, Hafner Publishing Co., New York, 1961.

[3] Wald, A., Sequential Analysis, John Wiley and Sons, Inc., New York, 1948.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 106-9	2. GOVT ACCESSION NO.: AD-A084041
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): AN INVESTIGATION OF A STATISTICAL PROCEDURE FOR MONITORING TWO-SAMPLE LIFE TESTS	5. TYPE OF REPORT & PERIOD COVERED: Technical Report
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Leslie A. Kalish and Dennis E. Smith	8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-1054
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Desmatics, Inc. P. O. Box 618 State College, PA 16801	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 042-334
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research Arlington, VA 22217	12. REPORT DATE: April 1980
	13. NUMBER OF PAGES: 19
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this report is unlimited.

17. DISTRIBUTION STATEMEN (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Life Tests
Repeated Testing
Staggered Entry
Power
Expected Test Length

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A statistical procedure is described for comparing two lifetime distributions when the data is reviewed repeatedly over time. The procedure provides the capability of early decision while maintaining both a fixed significance level and a fixed maximum length for the entire experiment. The effects which staggered entry, number of looks at the data and maximum test length have on power and expected test length are discussed. An application is made to a bearing fatigue-life test.

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)